idw - Informationsdienst Wissenschaft
Nachrichten, Termine, Experten

Pressemitteilung

CISPA Helmholtz Center for Information Security
Felix Koltermann

29.11.2024
http://idw-online.de/de/news843851

Forschungsergebnisse, Forschungsprojekte
Informationstechnik
überregional

Untersuchung von Webcrawlern legt Defizite offen

Der CISPA-Forscher Aleksei Stafeev legt zum ersten Mal eine Studie vor, in der das Wissen über Tools zur automatisierten Analyse von Websites, sogenannte Webcrawler, im Bereich der Web-Sicherheitsmessung systematisiert wird. Dafür untersuchte er hunderte Paper, die in den letzten 12 Jahren bei den wichtigsten internationalen Konferenzen publiziert wurden. Es zeigte sich, dass viele Paper die Crawler nur unzureichend beschreiben und dass randomisierte Algorithmen bei der Navigation der Crawler auf den Websites am besten abschneiden. Das Paper entstand im Rahmen des Projekts TESTABLE von CISPA-Faculty Dr. Giancarlo Pellegrino.

Studien zur Messung der Websicherheit etwa in Bezug auf die Umsetzung von Datenschutzmaßnahmen oder der Sicherheit von Websites erfreuen sich in der Cybersicherheitsforschung großer Beliebtheit. Für deren Umsetzung sind Crawler das Mittel der Wahl. „Ziel von Crawlern ist es, die Datenerfassung auf einer Website zu automatisieren“, erklärt CISPA-Forscher Aleksei Stafeev. Ihnen zu Grunde liegt ein Algorithmus der steuert, wie der Crawler automatisiert über eine Website läuft, verschiedene Seiten besucht und von diesen Daten sammelt. „Aber Webcrawling ist nicht so einfach, wie es klingt”, so Stafeev weiter. „Theoretisch besuchen die Tools einfach nur Websites. Aber in Wirklichkeit ist das Internet sehr komplex: Auf jeder Website gibt es eine Vielzahl verschiedener Schaltflächen und jede davon führt möglicherweise zu einer anderen Seite. Man hat ein exponentielles Wachstum verschiedener Seiten und muss herausfinden, welche man tatsächlich besuchen muss, um die für die eigene Forschungsfrage relevanten Daten zu erhalten.“ Trotz der großen Bedeutung von Webcrawlern wurde deren Leistung bisher nur sehr begrenzt untersucht. Diese Lücke schließt Stafeev nun mit seiner Studie.

Dabei ist der CISPA-Forscher zweischrittig vorgegangen. „Zunächst haben wir eine Übersicht über die aktuellen Arbeiten zu Web-Messungen durchgeführt, die Crawler verwenden“, erzählt Stafeev. Ergebnis war ein Datenkorpus von 407 Papern, die zwischen 2010 und 2022 publiziert worden waren. „Wir haben versucht, daraus die Informationen zu extrahieren, welche Crawler wie verwendet werden um ein allgemeines Bild davon bekommen, was bei Web-Messungen verwendet wird“, so der CISPA-Forscher. Für den zweiten Teil richtete Stafeev den Blick auf Paper der letzten drei Jahre, die neue Crawler vorschlagen. „Wir haben die Crawler im Hinblick darauf bewertet, welche Daten sie für den Zweck der Web-Sicherheitsmessung sammeln“, führt Stafeev weiter aus. Um die Crawler hinsichtlich der Code-Abdeckung, der Quellen-Abdeckung und der JavaScript-Sammlung untersuchen zu können, entwickelte Stafeev ein experimentelles Setup namens Arachnarium.

Unzureichende Beschreibungen und das Randomisierungs-Paradox

Eines der zentralen Ergebnisse des ersten Teils der Studie war, dass in den meisten Papern nur unzureichende Beschreibungen der Webcrawler zu finden waren. „Es war wirklich schwierig, die Informationen darüber, welche Technologie sie zum Crawlen verwenden und welche Techniken sie einsetzen, zu extrahieren und zu verstehen. Und es gab meist nicht genügend Details zu verwendeten Codes und Algorithmen. Oft hieß es nur ‚Wir verwenden Crawling‘ und das war's. Dass wir das als Community besser machen können, indem wir mehr Informationen über die von uns verwendeten Crawler und deren Konfiguration bereitstellen, war eine der wichtigsten Erkenntnisse.“ Wichtig ist dies vor

allem, um die Reproduzierbarkeit von Studien garantieren zu können, was ein zentrales Kriterium wissenschaftlicher Qualität darstellt.

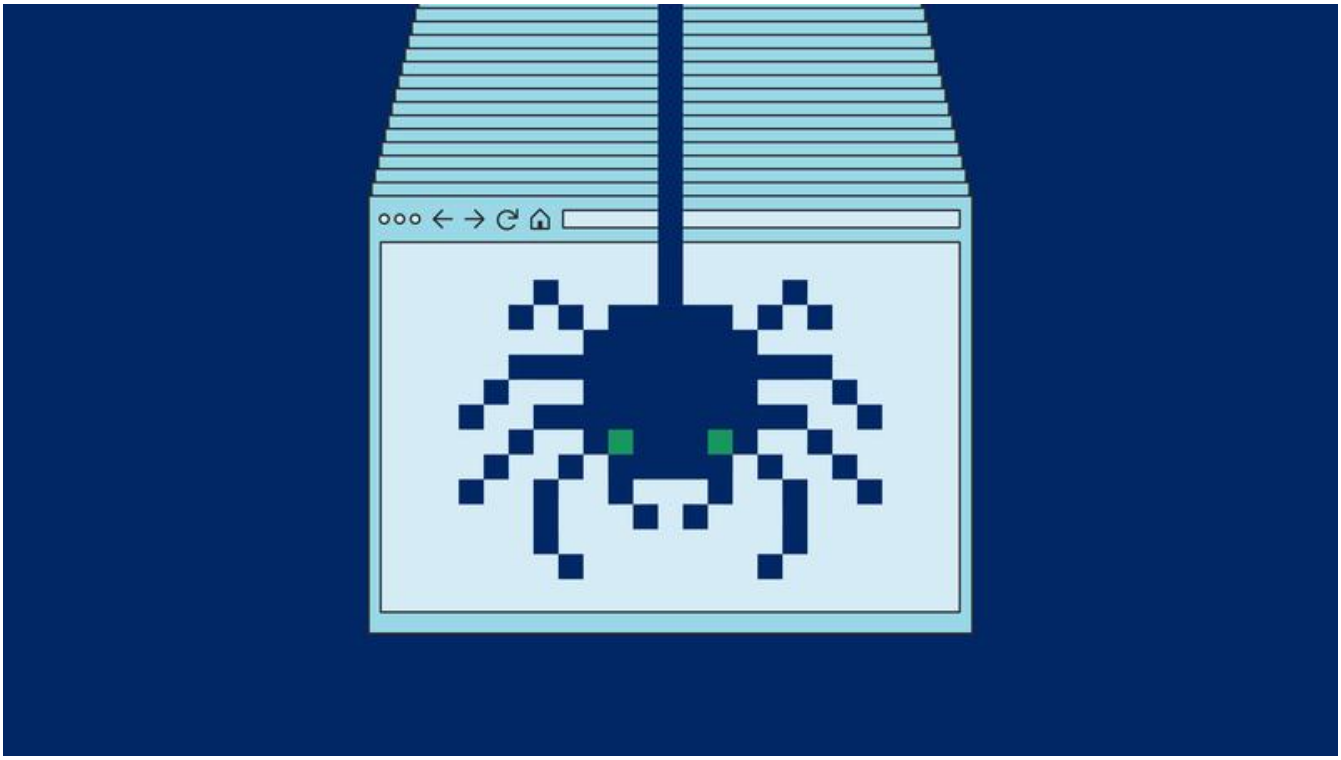
Ein erstaunliches Ergebnis förderte auch der zweite Teil der Studie zu Tage. „Nach unseren Daten scheinen Webcrawler, die randomisierte Algorithmen nutzen, am besten abzuschneiden“, erklärt Stafeev. „Das ist eigentlich ziemlich überraschend, bedeutet es doch, dass wir, egal was wir an Navigationsstrategien entwickelt haben, immer noch keine bessere Lösung gefunden haben, als einfach nur zufällig auf Dinge zu klicken.“ Der CISPA-Forscher testete Crawler mit verschiedenen Metriken. Dabei hat er festgestellt, dass es bei all diesen drei Metriken keinen einzigen Gewinner unter den Crawlern gibt. „Wir können also keine einheitliche Empfehlung für alle geben, die besagt: ‚Jeder sollte diesen Crawler verwenden‘ “, so der CISPA-Forscher weiter. Es hängt also entscheidend vom Kontext und der genauen Zielsetzung ab, welcher Crawler passend ist.

Take-Aways und der weitere Umgang mit den Forschungsdaten

Um die Studie umsetzen zu können, hat Stafeev einen riesigen Datensatz erstellt. „Wir glauben, dass wir noch viel mehr daraus lernen können“, erzählt er. „Und es wäre wirklich schön, wenn andere mehr Erkenntnisse aus den von uns gesammelten Daten gewinnen könnten.“ Aus diesem Grund hat Stafeev den kompletten Datensatz online frei zugänglich gemacht. Er selbst will sich in Zukunft wieder seiner eigentlichen Leidenschaft widmen: der Entwicklung neuer Crawler. Denn ursprünglich hatte Stafeev gar nicht geplant, eine so große Studie durchzuführen. Er wollte eigentlich nur seinen eigenen Crawler verbessern und sich dafür anschauen, wie andere mit dem Problem umgegangen waren. „Die Systematisierung von Wissen, wie sie dieser Studie zu Grunde liegt, ist ein ziemliche großes Unterfangen“, erzählt er. „Aber ich habe bei diesem Projekt extrem viel gelernt, wie man solche Experimente durchführt und mit so großen Datensätzen arbeitet. Dieses Wissen werde ich mir bei meiner künftigen Arbeit zunutze machen“, so der CISPA-Forscher abschließend.

Originalpublikation:

Stafeev, Aleksei; Pellegrino, Giancarlo (2024). SoK: State of the Krawlers - Evaluating the Effectiveness of Crawling Algorithms for Web Security Measurements. CISPA. Conference contribution.
https://doi.org/10.60882/cispa.25381438.v1



Visualisierung zum Paper "SoK: State of the Krawlers - Evaluating the Effectiveness of Crawling Algorithms for Web Security Measurements"
CISPA